PySpark Assignment

Saurav Sengupta (ss4yd)

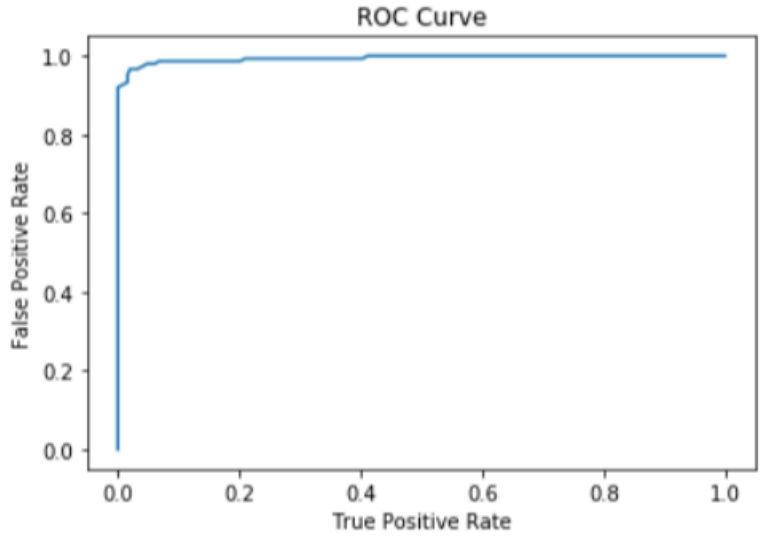
Motivation

- The data is related to Breast Cancer with the tumor size at various points in the patients lifetime.
- Aim to find relationship between tumor size and malignancy.
- Used PySpark libraries to build logistic regression and random forest models.

Code Snippet

- Used the summary function of the LogisticRegressionModel object to get summary.
- Then used matplotlib to plot the ROC curves.

In [37]:
```
roc = lrModel.summary.roc.toPandas()
plt.plot(roc['FPR'],roc['TPR'])
plt.ylabel('False Positive Rate')
plt.xlabel('True Positive Rate')
plt.title('ROC Curve for Logistic Regression')
plt.show()
print('Training set areaUnderROC: ' + str(lrModel.summary.areaUnderROC))
```




```
Training set areaUnderROC: 0.9942741935483873
```

Visualizations